

Paper 083-2007

Datamining to Investigate the Prescribing of Medications Longitudinally

Mussie Tesfamicael, University of Louisville, Louisville, KY

ABSTRACT

Forecasting is an important topic on corporate agendas to keep financial performance on track at a time when revenues are less predictable. In this paper, high performance forecasting was applied to healthcare data. SAS/STAT software facilitates the data processing and forecasting procedures to provide useful information that can help improve patient care and lower costs. The dataset was obtained from the medical expenditure panel survey (MEPS) by merging medications prescribed for the years 1996-2004. The variables used include the medication start date, medication quantity, patient insurance payment, and total cost of medication. SAS High Performance Forecasting (**PROC HPFDIAGNOSE & PROC HPFENGINE**) was used to investigate trends in medication prescription, usage, and cost. The medications were collected at a series of irregular intervals; we defined a monthly sum. PROC HPFENGINE automatically selects the IDM (Intermittent demand model) if most of the observations in the data are missing; as a result, either average or interval models are fitted. Cipro, Levaquin, Keflex and Tequin are the broad-spectrum antibiotics that will be demonstrated as examples of medication investigations. ARIMAX models, UCM (unobserved component model) and ESM (exponential smoothing models) will be fitted to the antibiotics datasets to forecast the total cost, private insurance payment, quantity of prescribed medication and the number of prescriptions. Text Miner in SAS Enterprise Miner was used to cluster the severity of the patient diseases based on the ICD9 codes (developed by the World Health Organization) listed for each patient. The clusters were labeled using domain knowledge. Finally, Private insurance payment and Total payment of the medications were compared for several clusters. Kernel density estimation (PROC KDE) was used to examine differences in the clusters to compare the severity of the disease and cost.

Key words: High Profile Forecasting, ARIMAX, ESM, UCM, IDM and Text Mining.

INTRODUCTION

The purpose of this paper is to examine the use of transactional time series data in a healthcare industry in order to predict the trends in prescription practices. A statistical time series is an indexed set of numbers, where the time series are equally spaced. A time series is equally spaced if any two consecutive indices have the same time difference. If the data are collected at specific time intervals, say at each hour, day or month, time series analysis methods can be applied. However most data are collected at irregular intervals of time, called transactional data. A problem arises when the time interval for the data points are irregular intervals. The data to be analyzed in this paper were collected from the medical expenditure panel survey (MEPS) [1]. The MEPS contains new and extensive data on the use of health services and healthcare in the United States. MEPS provide nationally representative estimates of healthcare use, expenditures, sources of payment, and insurance coverage for the U.S. civilian non-institutionalized population. The MEPS consists of three surveys. The Household Component (HC) is the core survey and forms the basis for the Medical Provider Component (MPC) and part of the Insurance Component (IC). Together, these surveys yield comprehensive data that provide national estimates of the level and distribution of health care use and expenditures, support health services research, and can be used to assess health care policy implications. The quality of healthcare providers can be ranked by the services they provide to the public. When we consider quality in the healthcare industry, we also have to consider the severity of a patient's condition. Sicker patients will be more at risk for complications and mortality. A healthcare provider that admits a higher proportion of severe patients will have a lower rate of favorable outcomes compared to a healthcare provider that admits only low risk patients. Patient severity is determined by the patient's entire health status, which is determined by patient illness. The patient condition is defined by a list of ICD9 (ICD10) codes developed by the World Health Organization [2].

The rising cost of medications is an important concern to health care providers and to society. Ewing, Piette, and Payne (2003, 2004) studied the time series properties of several medical net discount rates. In order to calculate the present value of future medical costs, the practitioner needs a forecast of the appropriate medical net discount rate. For many years, proper application of antibiotics has been difficult to regulate and to control. Antibiotic costs have increased dramatically over the years with an overall trend to prescribe expensive broad spectrum rather than narrow-spectrum antibiotics. [6, 7]

The term trend of antibiotic cost is an important aspect to health care providers and insurance companies. The trend analysis depends largely on an initial explanatory analysis of the data and in identifying the appropriate model orders to predict the trend. The best approach currently available to model trend is to eliminate the trend by differencing and data correction and to find an appropriate stationary model for the differenced series. Trend modeling requires finding the appropriate order of differencing, to correct the data for missing values and to identify the appropriate order of

stationary models for the differenced and corrected data. If the data to be analyzed are non-stationary, differencing is done until the data are stationary, and then an appropriate model is built.

This study explores the prescription and cost trend of the antibiotics: Cipro, Keflex, Levaquin, and Tequin. An appropriate model was built by using PROC HPFDIAGNOSE and PROC HPFENGINE. Several models from SASHELP.HPFDFTL were selected in a MODELREPOSITORY (temporary holder of models) of PROC HPFDIAGNOSE, and then PROC HPFENGINE selected an appropriate model. PROC HPFENGINE analyzed the data and, when most observations were found missing, an intermittent demand model (IDM) was selected; otherwise models such as ESM(exponential smoothing model), UCM(Unobserved components model) or ARIMA models were fitted. PROC HPFDIAGNOSE also identified the autocorrelation function (ACF), partial autocorrelation function (PACF), and the inverse autocorrelation function (IACF), the type of differencing needed, and residual analysis. A graphical output of the result will be displayed to show how good the model fits the data.

Text Miner in Enterprise Miner was used to perform clustering of the data. With domain knowledge, a name was given to each cluster in order to obtain a meaningful conclusion. A PROC KDE was used to compare the severity of the diseases and the cost of the antibiotics.

DATA PROCESSING

When analyzing transactional data with time series methods, the variable of interest is the time point at which the event happens; in this case, the time a patient started taking medications. Data exploration was done by concatenating the prescribed medications for the years 1996-2004 (each year is contained in a separate data file). The data for each year was in SAS transport format, so the following code is written to read the file in SAS,

SAS CODE1

```
LIBNAME PUFLIB V8 'C: \MEPS',
FILENAME IN1 'C: \MEPS\H67A.SSP';
PROC XCOPY IN=IN1 OUT=PUFLIB IMPORT;
RUN;
```

For each year, the data were filtered by Cipro, Keflex, Levaquin and Tequin prescriptions. Merging the columns day, month and year of medication prescribed created a transactional time variable called StartMedDate. The time series date variable should have a unique value, but the created transactional time variable has duplicate values as two different households report that they buy the medication at the same date of the month. The time variable collected is irregular, and time series methods won't analyze irregular data points, so we have accumulated the information by total transactions for a given month. The SAS code, **PROC HPF**, has a statement called **ACCUMULATE=OPTION** and **LEAD=0**, so the user can select the accumulation point for the data; in this paper, we have accumulated the data monthly. The other problem with this dataset was missing values for the day, month or year of prescription; this usually happens when a patient doesn't remember the date they started taking medications, so this leads to a missing value. However, since we are taking the accumulation point to be monthly, we have assigned a value of 01 for a missing day. If a month or year of the medication started is missing, the values can't be imputed, as this can distort the trend of the time series. The SAS code StartMedDate= MDY (Var1, Var2, Var3) forms an appropriate date variable, where Var1=month, Var2=day, and Var3=year. If a series of data has largely zero or missing values, it is better predicted by intermittent demand models (IDM). The PROC HPFDIAGNOSE first diagnoses whether the series is intermittent, in which case an IDM model is built; other wise, other models are built. Intermittent time series commonly occur in Internet, inventory, sales, survey, and other situations where the demand for a particular item is occasional.

In this paper, graphics, clusters of ICD9 (ICD10) codes, and model testing are used as hypothesis testing and processing methodologies. The graphics include density plots on which the Kernel density estimation (KDE) was used to determine the distribution of the data points as well as the total payment and Insurance payment that were forecasted and were compared to the actual fit. Autocorrelation plots (ACF, IACF & PACF) were used to check whether the data points were autocorrelated. For the data with most observations missing, an IDM model was built.

MODELS USED FOR ANALYSIS

HPF DIAGNOSE

The HPFDIAGNOSE procedure automatically diagnoses the statistical characteristics of time series and identifies appropriate models. The models that HPFDIAGNOSE considers for each time series include ARIMAX (Autoregressive Moving Average with integrator in the noise model and exogenous input (predictor variable)

variables, ESM (Exponential Smoothing), and UCM (Unobserved Components models). Log transformation and stationary tests are automatically performed. The ARIMAX model diagnostics find the AR and MA orders, detect outliers, and select the best input variables. The Unobserved Components Model procedure analyzes and forecasts equally spaced univariate time series data using the Unobserved Components Models (UCM). The UCMs are also called Structural Models in the time series literature. A UCM decomposes the response series into components such as trend, seasonals, cycles, and the regression effects due to predictor series. The components in the model are supposed to capture the salient features of the series that are useful in explaining and predicting its behavior.

ARIMA MODELS

An Auto regressive moving average model (ARIMA) is a combination of Autoregressive (AR) and Moving Average (MA) parts. A model that is purely autoregressive with autocorrelations means that the current value y_t is dependent upon a specific number of previous values with an autoregression of size p , written as

$y_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t, t = 1, \dots, T$ $\varepsilon_t \sim N(0, \sigma_t^2)$. We can estimate the number of lags that are used in the model based upon the autocorrelations. The moving average component of size q expresses the current value y_t in terms of future errors, and mathematically, it is written as $y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, t = 1, \dots, T$.

If a trend exists or the model is not stationary, a first, second or higher order differencing is taken. Therefore, a new model can be defined as $W_t = y_t - y_{t-1}$; this is what is called differencing. Once we estimate $W_t, t = 1, \dots, T$, we estimate y_t as $y_t = W_t + y_{t-1}$. Therefore, the general autoregressive-moving average process is written as

$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$. The ARIMA model is identified as order (p, d, q) , which estimates both the autocorrelation and trend terms. The ARIMA model with a seasonality d is written as

$ARIMA(p, d, q) \times (P, D, Q)$, where P =number of seasonal autoregressive terms, D =number of seasonal differences, and Q =number of seasonal moving average terms. In the seasonal part of the model, all of these factors operate across multiples of lag s (number of periods in a season). If the seasonality changes on a yearly basis, then the value of s is equal to 12, whereas if the seasonality changes on a quarterly basis, then the value of s is equal to 4.

INTERMITTENT DEMAND MODEL

Intermittent time series have a large number of values that are zero. Typically, when the value of the series associated with a particular time period is nonzero, *demand* occurs; and, when the value is zero (or missing), *no demand* occurs. Since it is entirely possible that the number of time periods for which *no demand* occurs is large, many of the series values will be zero. Typical time series models (for example, smoothing models) are inadequate in the case of intermittent time series because many of the series values are zero. Since these models are based on weighted-summations of past values, they bias forecasts away from zero. Unlike the smoothing models that provide forecasts for future time periods, intermittent forecasting models provide recommended *stocking levels* or *estimated demand per period* that are used to satisfy future demand. An intermittent time series (demand series) can be decomposed into two components: a demand interval series and a demand size series. Both of these component series are indexed based on when a *demand* occurred (demand index) rather than each time period (time index). The demand interval series is constructed based on the number of time periods between *demands*. The demand size series is constructed based on the size (or value) of the *demands* excluding zero (or base) demand values. Using these two component series, the average demand series is computed by dividing the size component values by the interval component values. Given a time series Y_t , for $t = 1, \dots, T$, where t is the time index, suppose that there are N nonzero demands occurring at times $t = t_i$, where $t_{i-1} < t_i$, for $t = 1, \dots, N$. The time series is dissected into the demand interval series and the demand size series as follows: Demand Interval Series $q_i = t_i - t_{i-1}$ for

$t = 2, \dots, N$, Demand Size Series $d_i = Y_1 - base$ for $i = 1, \dots, N$ and Average Demand Series $a_i = \frac{d_i}{q_i}$ for $i = 2, \dots, N$ [3].

EXPONENTIAL SMOOTHING MODELS

The ESM statement in the HPFDIAGNOSE procedure specifies the exponential smoothing model and the options used to control the parameter estimation process for the model. The model for Total payment and Insurance payment for Keflex is built with the Exponential Smoothing Model. The exponential smoothing model, as its name implies, builds a model using the exponential smoothing function.

Plotting of the original variable of interest versus time gives a general overview of the characteristics of the time series. From some plots, we can easily observe seasonality, trend or irregular component. However, descriptive analysis and trend modeling need to be checked.

The cost of antibiotics y_t is correlated with the previous cost y_{t-1} . This correlation between series points needs to be checked before a suitable model is fit. In autocorrelated time series, negative deviations from the mean are followed by positive deviations and positive deviations by negative, whereas positive deviations from the mean are followed by positive deviations and negative deviations by negative deviations. The SAS PROC HPFDIAGNOSE and PROC HPFENGINE procedures plot the prediction error autocorrelation function (ACF), prediction error partial autocorrelation function (PACF), and prediction error Inverse autocorrelation function (IACF). Significant spikes in the ACF, PACF, and IACF plots at specified frequencies indicate the presence of autocorrelation at the frequencies (Figures 3- 5). Estimators of the inverse autocorrelation function in SAS PROC ARIMA are computed by fitting a long autoregressive approximation to the data, moving the autoregressive coefficients over to the moving average model, then algebraically calculating the autocorrelations for it. This method of calculation, when applied to the first differences of a stationary series, yields a slowly decaying estimated inverse autocorrelation. One has to be careful not to over-difference a series. The plots of ACF, IACF and PACF for the antibiotics data did not violate this assumption [8].

AUTOCORRELATION FUNCTION PLOT FOR TOTAL PAYMENTS (RXXPX)

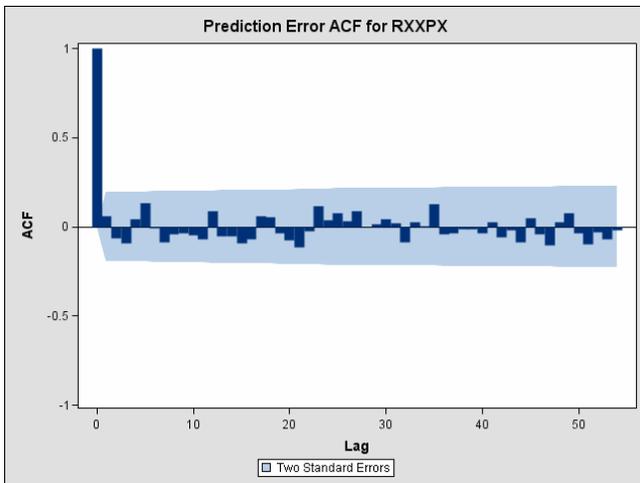


Figure 1. ACF plot of total payment of Cipro

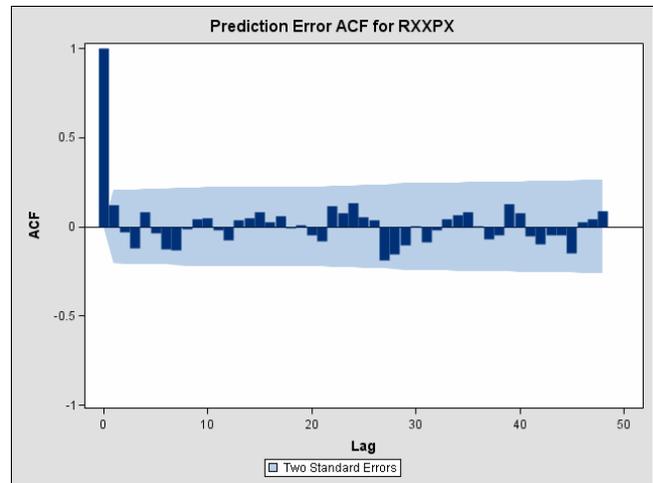


Figure 2. ACF plot of total payment of Keflex

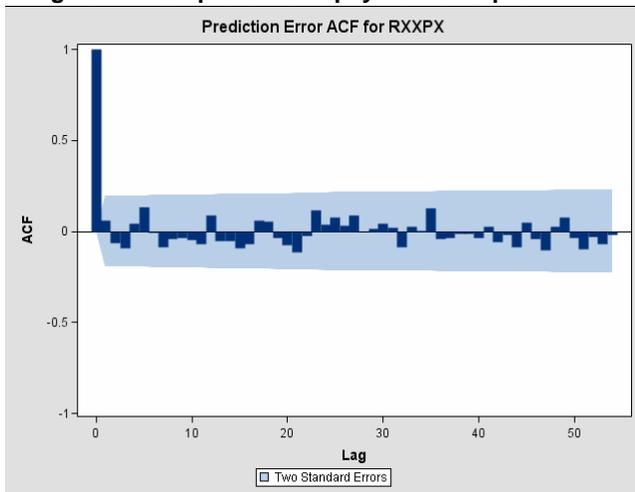


Figure 3. ACF plot of total payment of Levaquin

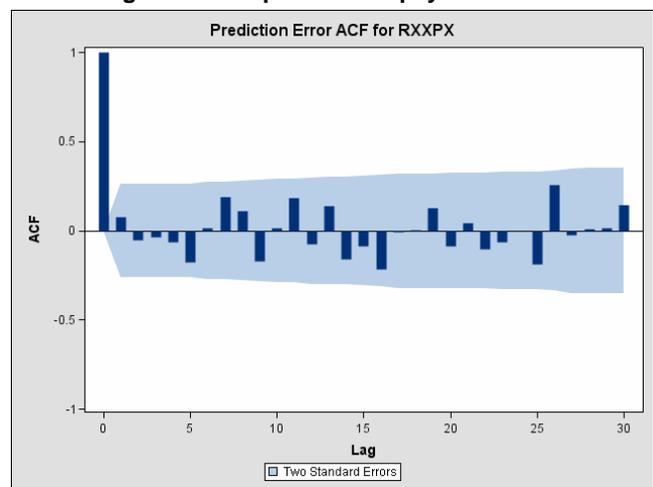


Figure 4. ACF plot of total payment of Tequin

The prediction error plot for autocorrelation function (Figures 1-4) all lie within two standard errors; this is an indication that the model is a good fit. These error plots are for the total payment made for the antibiotics CIPRO, KEFLEX, LEVAQUIN and TEQUIN.

Next we display the partial autocorrelation plots; these plots provide similar information on how well the model built is fitting the data.

PARTIAL AUTOCORRELATION FUNCTION PLOT FOR TOTAL PAYMENTS (RXXPX)

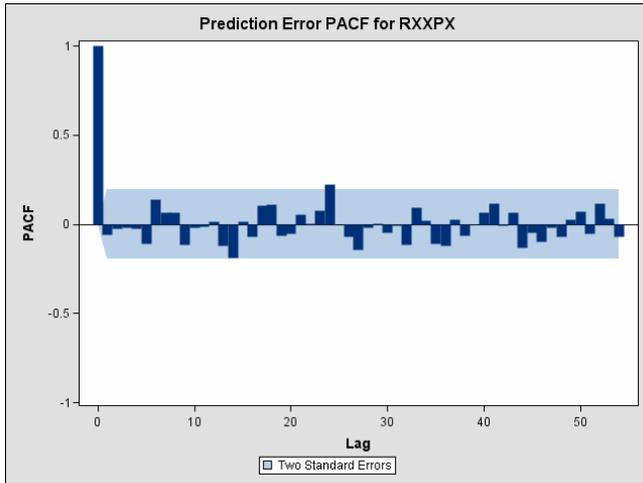


Figure 5. PACF plot of total payments of Cipro

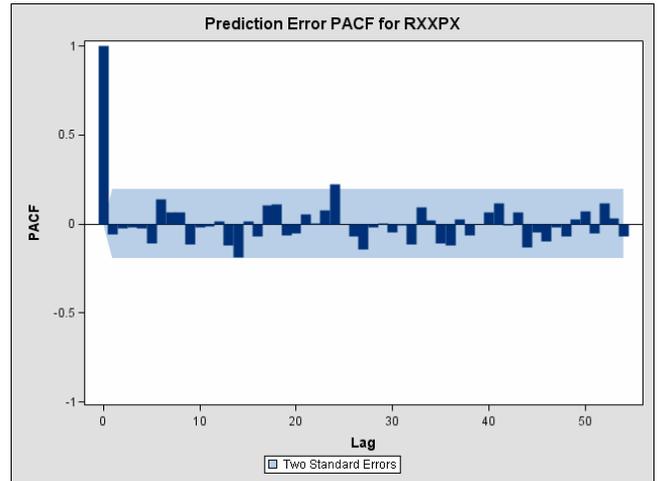


Figure 6. PACF plot of total payment of Keflex

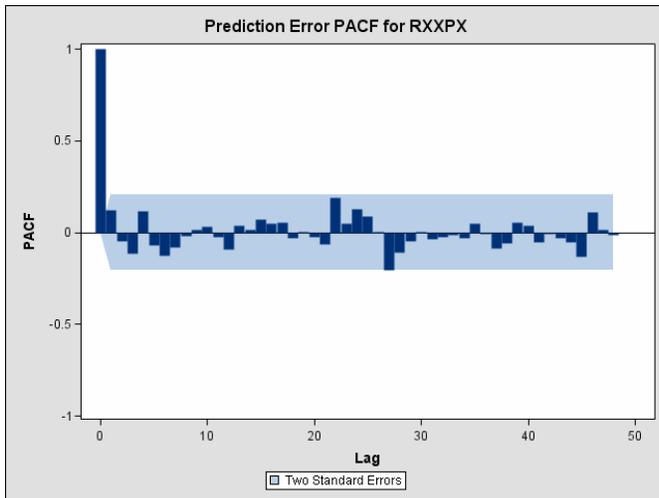


Figure 7. PACF plot of Total payment of Levaquin

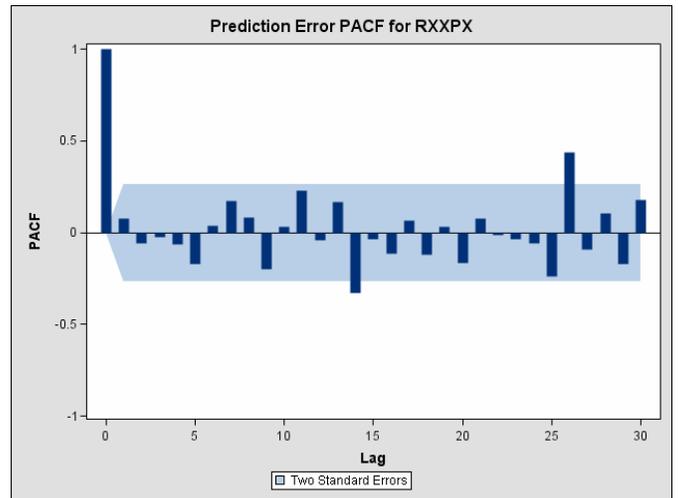


Figure 8. PACF plot of Total payment of Tequin

The error prediction for the partial autocorrelation function (Figures 5- 8) for the models built shows that except for TEQUIN, which has two spikes outside the 2 standard errors, the remaining lags lie within the confidence error. The graphs here reveal that the models built fit the data well.

Next we display the autocorrelation function plots and partial inverse autocorrelation plots for model prediction error for Private insurance payments made for the antibiotics CIPRO, KEFLEX, LEVAQUIN and TEQUIN. These plots give an idea on the fit of the models.

AUTOCORRELATION FUNCTION PLOT FOR PRIVATE INSURANCE PAYMENT (RXPVX)

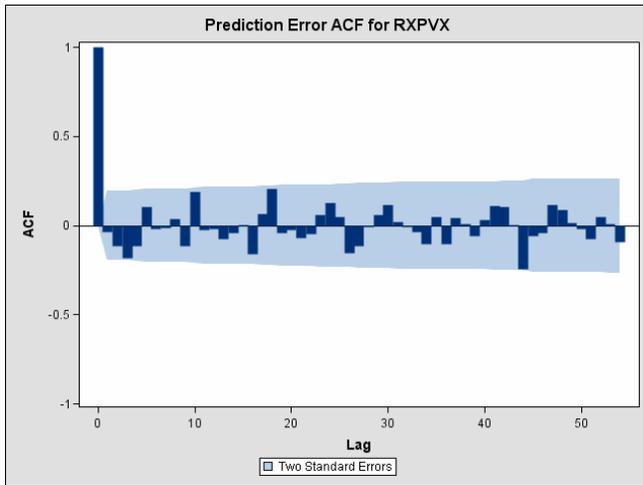


Figure 9. ACF plot of RXPVX payment of Cipro

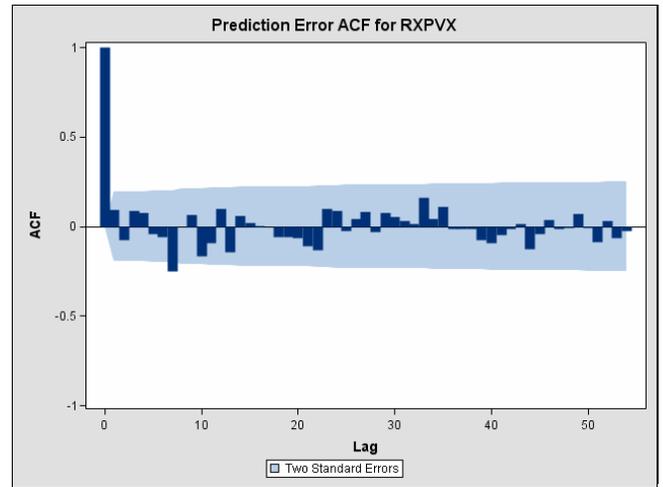


Figure 10. ACF plot of RXPVX of Keflex

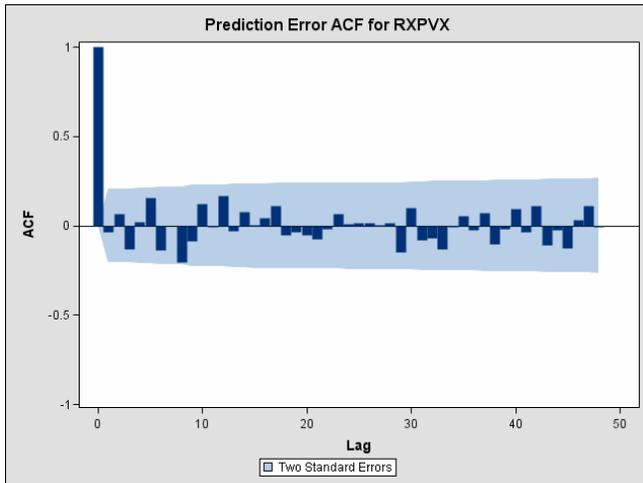


Figure 11. ACF plot of RXPVX of Levaquin

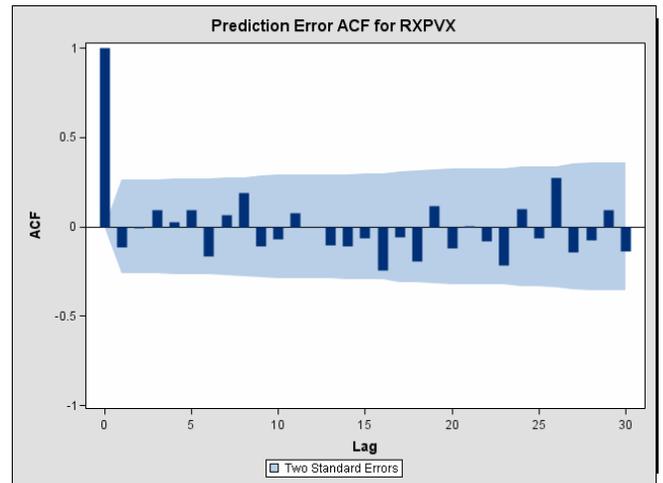


Figure 12. ACF plot of RXPVX of Tequin

The prediction error plot for the autocorrelation function (Figures 9-12) all lie with in two standard errors; a good fit. These error plots are for the Private Insurance payment s made for the antibiotics CIPRO, KEFLEX, LEVAQUIN and TEQUIN.

PARTIAL AUTOCORRELATION FUNCTION PLOT FOR PRIVATE INSURANCE PAYMENT (RXPVX)

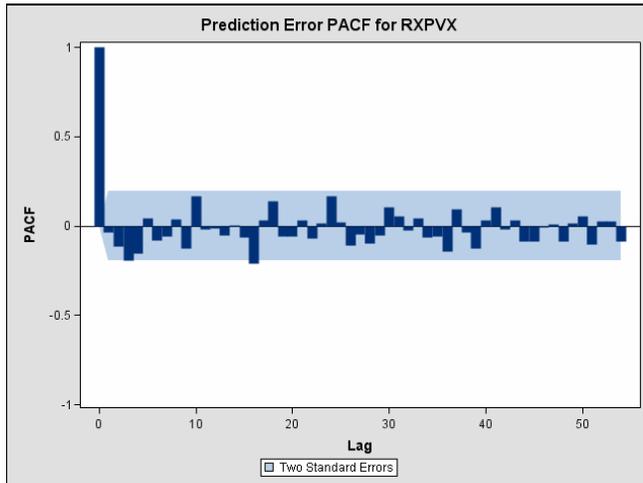


Figure 13 PACF plot of RXPVX FOR Cipro

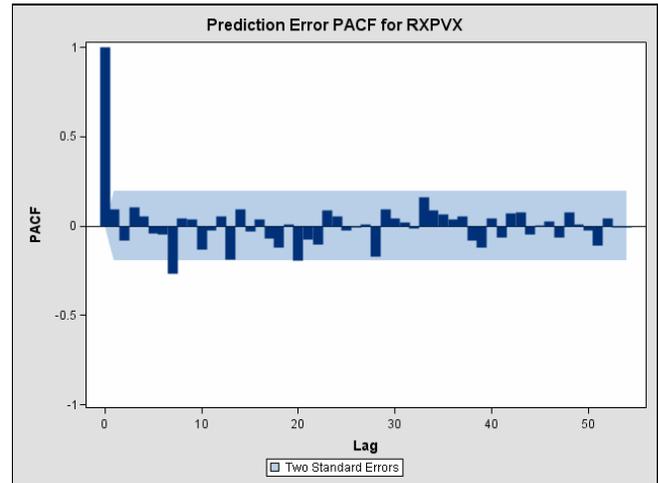


Figure 14 PACF plot of RXPVX FOR Keflex

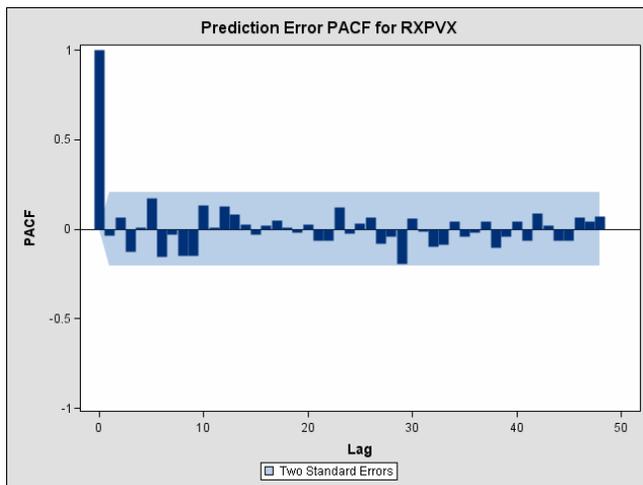


Figure 15 PACF plot of RXPVX FOR Levaquin

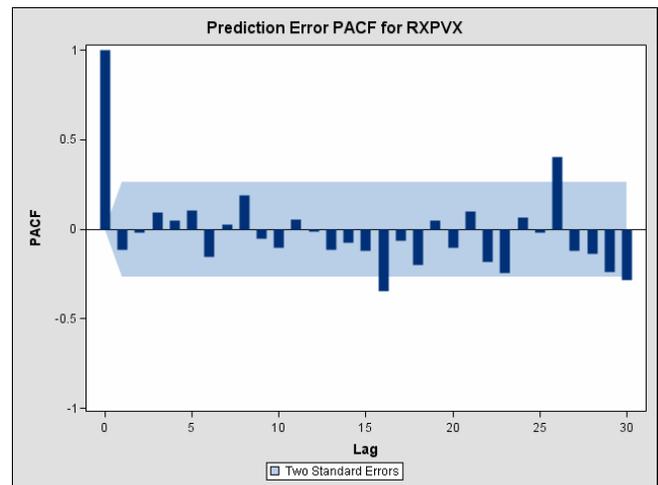


Figure 16 PACF plot of RXPVX FOR Tequin

The prediction error plot for partial autocorrelation function (Figures 13-16) all lie within two standard errors except Figure 16 there is a spike at lag 17 and lag 28. These error plots are for the Private Insurance payments made for the antibiotics Cipro, Keflex, Levaquin and Tequin.

Outliers are an important issue to consider in statistics. The question is how to detect them, and once we know that there is an outlier, do we remove it or keep it? In time series forecasting, a data set with an outlier can have large error. In order to reduce the error and get a good model fit, we have to see if the difference between the predicted value from the actual value. If the actual value at time t , say y_t , is within two standard deviations of the mean, then the value is not an outlier, but if $2 \times SD < y_t \leq 3 \times SD$, then an intervention variable must be created to control the error. If the observation is $y_t \geq 4 \times SD$ then the observation would be taken as an outlier and can be eliminated from the data.

NON-STATIONARITY, DIFFERENCING and TREND

A time series has a stationary process if it satisfies these two conditions; 1) the series has a finite mean and finite variance that does not change with historical time, 2) the covariance between two random variables at t and $t+k$ is a function only of their relative lag, k ; not of the starting point. In other words, the covariance of this process is independent of historical time [9]. In practice, most time series are non-stationary and have a trend. Often the non-stationary component is of importance. However, much of the probability theory of time series is concerned with stationary time series, and most of the stochastic time series models require the data to be stationary [10]. If the time series is non-stationary, a differencing must be done. Differencing removes the trend and seasonality of the original time series data. The HPFDIAGNOSE and HPFENGINE procedures in SAS CODES2 automatically check if the series is non-stationary, in which case differencing is used and model diagnostics are computed as well.

```

PROC HPFDIAG DATA=SASUSER.CIPROTIMESERIES OUTEST=EST_THREE
BASENAME=AMXESM PRINT=ALL
MODELREPOSITORY=SASUSER.ANTIBIOTICS CRITERION=RMSE;
  ID STARTMEDDATE INTERVAL=MONTH;
  FORECAST RXXPX;
  ARIMAX PERROR=(12:24) P=(0:12) Q=(0:12) CRITERION=SBC METHOD=MINIC;
  ESM;
  UCM;
RUN;
ODS HTML;
ODS GRAPHICS ON;
PROC HPFENGINE DATA=SASUSER.CIPROTIMESERIES
MODELREPOSITORY=SASUSER.ANTIBIOTICS INEST=EST_THREE
GLOBALSELECTION=TSFSSELECT
PRINT=(SELECT ESTIMATES)
LEAD=12
OUTFOR=CIPROTOTAL
OUTEST=CIPROEST
OUTSTAT=CIPROSTAT PLOT=ALL;
FORECAST RXXPX;
  ID STARTMEDDATE INTERVAL=MONTH;
RUN;
ODS HTML CLOSE;
ODS GRAPHICS OFF;

```

SAS CODE2

The HPFDIANGNOSE procedure builds models from the selected list (ARIMAX, ESM and UCM). The MODELREPOSTORY is a ware-house for the models built,; the model estimates are put in OUTEST=. The HPFENGINE procedure builds the best model from the MODELREPOSTITORY and also uses some models from GLOBALSELECTION= where the new models built are stored in MODELREPOSITORY=SASUSER.ANTIBIOTICS. The statement OUTFOR= OUTEST= and OUTSTAT= put the values of the forecast, estimates and statistics respectively in datasets. The PERROR= specifies the range of the AR order for obtaining the series, P= specifies the AR order, Q= specifies the range of the MA order, CRITERION=SBC specifies that the Swartz Bayesian Criterion is selected and METHOD=MINIC (Minimum information Criterion).

ANALYSIS OF PRESCRIPTIONS 1996-2004

The main aim of this study is to build a model that predicts the Total Payments and Insurance Payments for the antibiotics CIPRO, KEFLEX, LEVAQUIN and TEQUIN. Knowing the trend of the increase in cost is helpful for healthcare industries as this can help them to question why the cost has increased in one antibiotic related to another. We even can question whether there was some kind of outbreak that makes a sudden shift in cost and quantity of prescriptions. High performance forecasting is very useful in healthcare industries as it estimates the trends in cost and quantity.

FORECASTED AND HISTORICAL PLOTS OF TOTAL PAYMENTS (RXXPX)

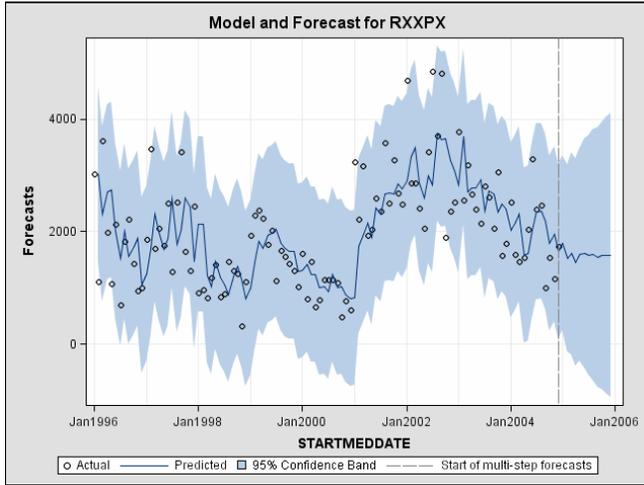


Figure 17. ARIMA: RXXPX-P=(1,2,3,4) D=(1) FOR Cipro

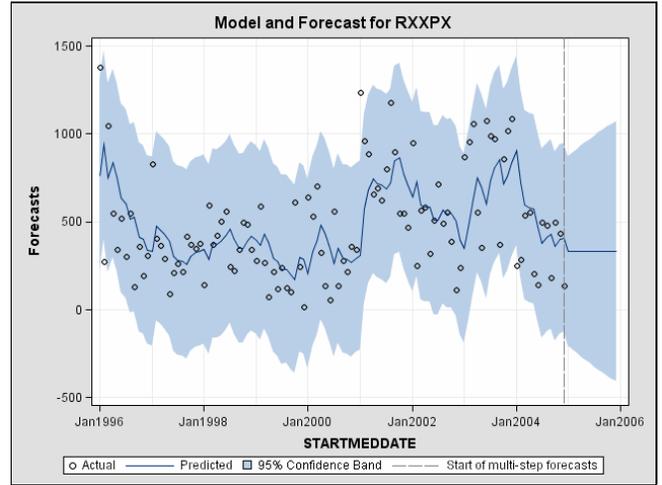


Figure 18. ARIMA: RXXPX-ESM FOR Keflex

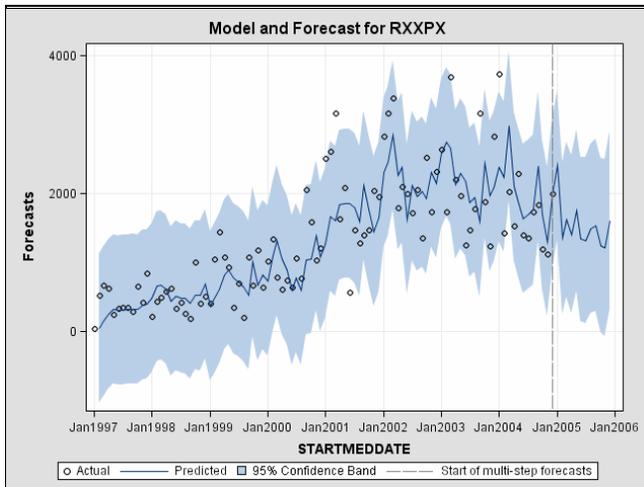


Figure 19. ARIMA: RXXPX-P=(1 2) D=(1) Q=(1) FOR Levaquin

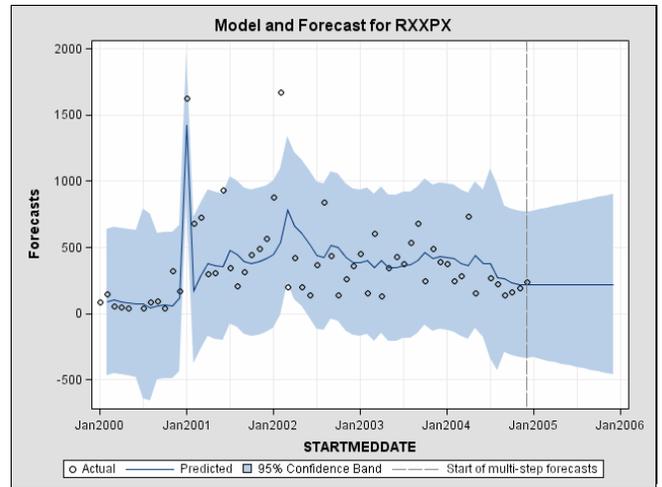


Figure 20. ARIMA: RXXPX-D=(1) Q=(1) Tequin

The total payments made on Levaquin are increasing (Fig 19) over the period of 1997-2005, with a peak at the beginning of winter. The total payment of CIPRO has an irregular trend until 2001; then suddenly there is an increase in total payment. For Keflex, the Total payment was decreasing until 2001 when there was an increase afterward. For TEQUIN, there was a linear increase in Total payment and the forecast becomes constant. There is a sudden shift in total payment for Cipro, so it would be good to study what happened at that incident.

FORECASTED AND HISTORICAL PLOTS OF PRIVATE INSURANCE PAYMENT (RXPVX)

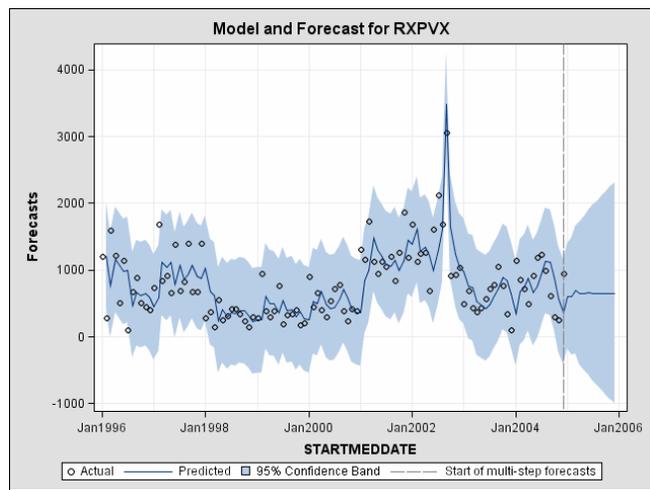


Figure 21. ARIMA: RXPVX~P=(1,2) D=(1) FOR Cipro

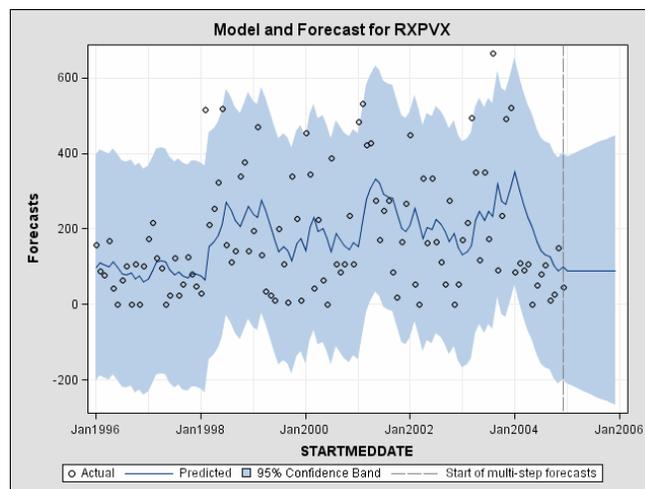


Figure 22. ARIMA: RXPVX~ESN for Keflex

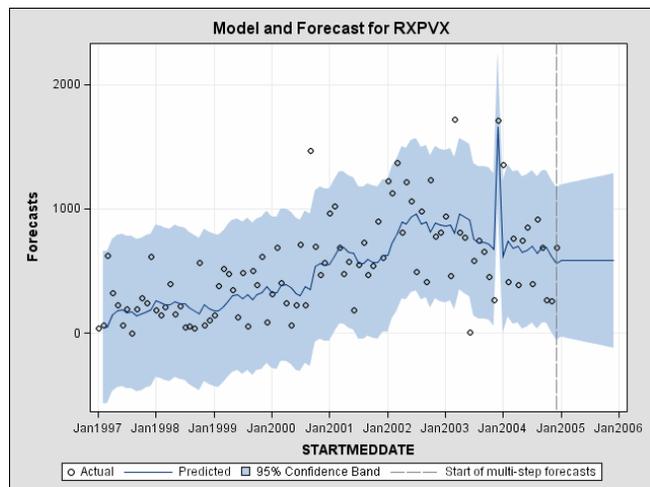


FIG 23. ARIMA: RXPVX~D=(1) Q=(1) FOR Levaquin

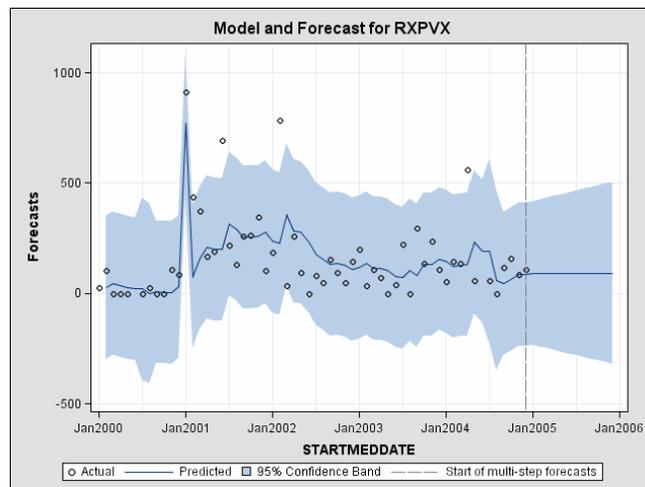


FIG 24. ARIMA: RXPVX~D=(1) Q=(1) FOR Tequin

The Private Insurance payment made for LEVAQUIN is increasing (Fig 23), with a shift at Jan 2004. Private Insurance payments made on TEQUIN were increasing until the middle of 2001 and then kept constant, with Private Insurance payment shifting at Jan 2001. The Private Insurance payment for Cipro is somewhat constant until Jan 2001, then starts increasing until mid of 2002, where there is a shift and then kept constant afterward. For Keflex there is a shift at Jan 2001, but it was not increasing.

Antibiotics Use and ICD9 Codes using Text Analysis

Patient severity is determined by the patient's entire health status, which is determined by patient illness. The codes are a series of 5-digit numbers with the first 3 digits representing the main condition and the last 2 digits representing specifics of the condition. For example, "599" represents other disorders of the urethra and urinary tract. In this paper,

all ICD9 codes will be treated as text rather than as categories. In this way, similarities between codes can be related to similarities in patient conditions, taking full advantage of the stemming properties contained within the codes.

The data must be partitioned routinely into three sets: training, testing and validation. The training set is used to define the model in Enterprise Miner. The validation set is used iteratively to ensure that the developed model fits the new dataset. The testing set is used to make a final comparison. The first step in analyzing text data is to define a term by document matrix. Each document forms a row of the matrix; each term forms a column. The resulting matrix will be extremely large, but very sparse with most of the cells containing zeros. The matrix can be compressed using the technique of singular value decomposition with the matrix restricted to a maximum of N dimensions. Singular value decomposition is based on the assignment of weights to each term in the dataset. Terms that are common and appear frequently, such as 'of', 'and', 'the' are given low or zero weight while terms that appear in only a handful of documents are given a high weight.

Clustering was performed using the expectation maximization algorithm. It is a relatively new, iterative clustering technique that works well with nominal data in comparison to the K-means and hierarchical methods that are more commonly used.

Table1. Text Clusters Defined by Expectation Maximization

Cluster Number	ICD-9 Codes	ICD-9 Risk Factors	Frequency	Label
1	522 462 005 487 780	diseases of pulp and periapical tissues acute pharyngitis Other food poisoning (bacterial) influenza Alteration of consciousness	20	Moderate risk factors
2	473 599 724 616 535	chronic sinusitis other disorders of urethra and urinary tract Inflammatory disease of cervix, vagina, and vulva gastritis and duodenitis gastritis and duodenitis	26	Urinary tract infection
3	595 682 V68 729 596	cystitis other tract of urinary disease Other cellulitis and abscess (INFECTION OF SKIN) Encounters for administrative purposes other disorders of soft tissues other disorders of bladder	17	Routine problems
4	518 463 V07 553	other disease of lung acute tonsillitis need for isolation and other prophylactic measures hernia of abdominal cavity	12	Mild risk factors
5	493 706 477 136 959	asthma Diseases of sebaceous glands allergic rhinitis Other and unspecified infectious and parasitic diseases injury, other and unspecified	18	Infectious disease
6	536 041 465 787 733	disorders of function of stomach Bacterial infection in conditions classified Acute upper respiratory infections of multiple or unspecified sites symptoms involving digestive system Other disorders of bone and cartilage	6	Upper respiratory infection

Table 1 gives the codes and their corresponding medical terms used to define the cluster labels. The text analysis was performed using the statistical software, SAS.

A weighted univariate kernel density estimate involves a variable X and a weight variable W . Let (X_i, W_i) , $i = 1, 2, \dots, n$ denote a sample of X and W of size n . The weighted kernel density estimate of $f(X)$, the density of

X , is as follows: $f(\hat{X}) = \frac{1}{\sum_{i=1}^n W_i} \sum_{i=1}^n W_i \varphi_h(X - X_i)$ where h is the bandwidth and

$\varphi_h(X) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{X^2}{2h^2}\right)$ is the standard normal density rescaled by the bandwidth.

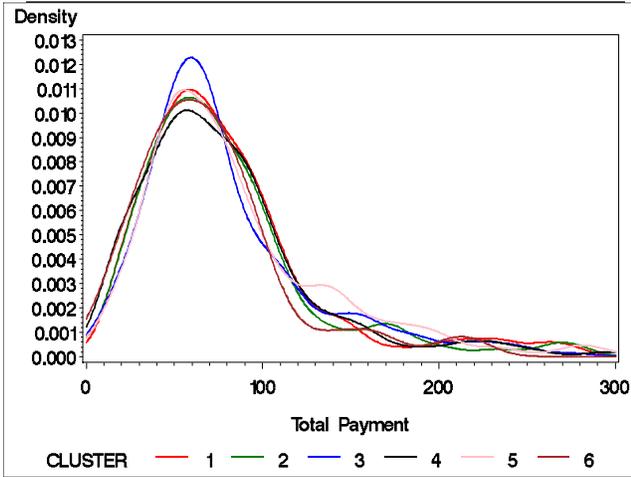


Figure 25. Distribution of Total Payments for Cipro

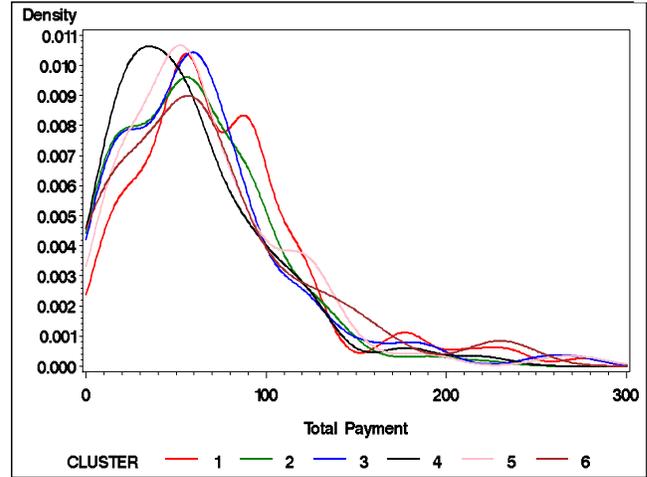


Figure 26. Distribution of Total Payments for Keflex

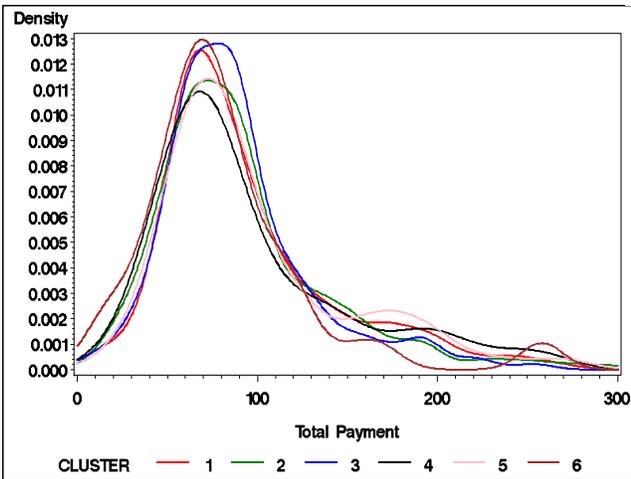


Figure 27. Distribution of Total Payments for Levaquin

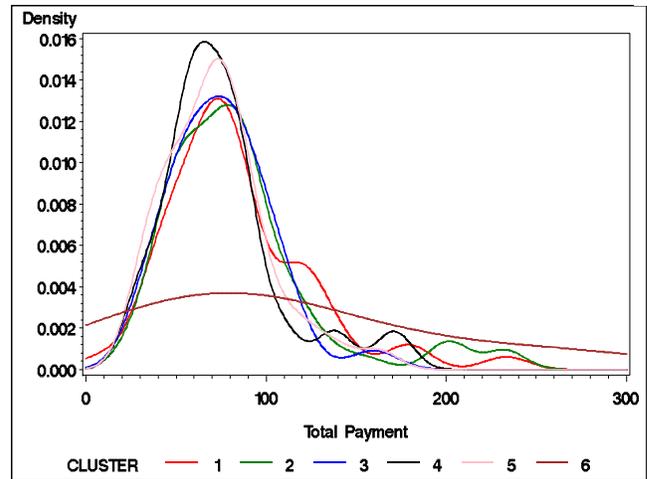


Figure 28. Distribution of Total Payments for Tequin

The clusters formed based on the severity of the patient condition were compared using Kernel density estimation of the private insurance payment for each antibiotic (Figures 25-28). For instance, for patients taking Levaquin in cluster 3 (Figure 27), there is a higher payment between 80-120 dollars. As the total payment exceeds 280 dollars, the distribution is relatively similar for each cluster.

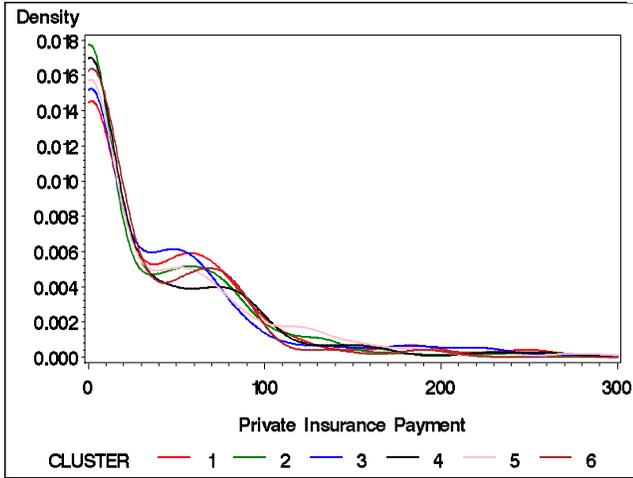


Fig 29. Distribution of Private Insurance for Cipro

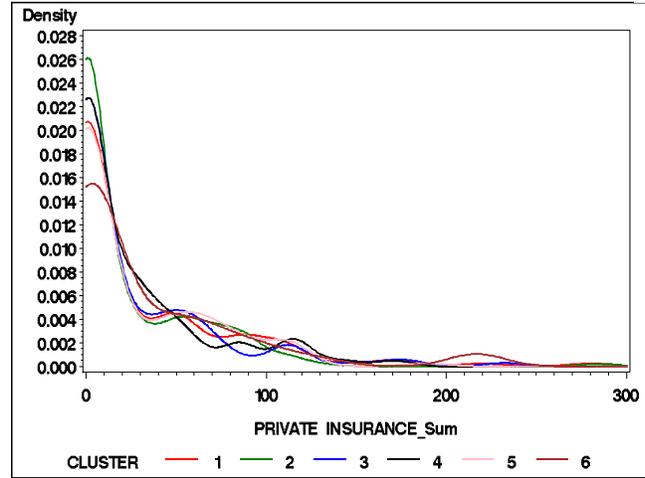


Fig 30. Distribution of Private Insurance for Keflex

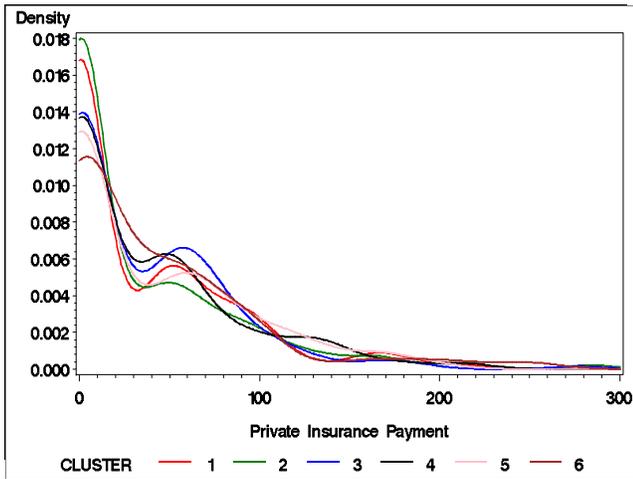


Fig 31. Distribution of Private Insurance for Levaquin

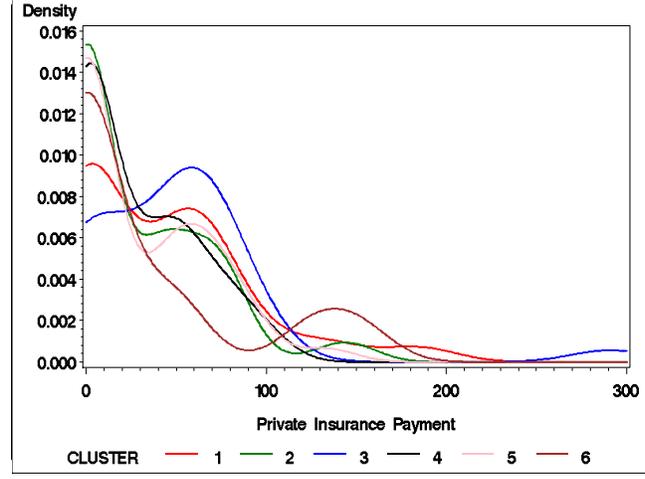
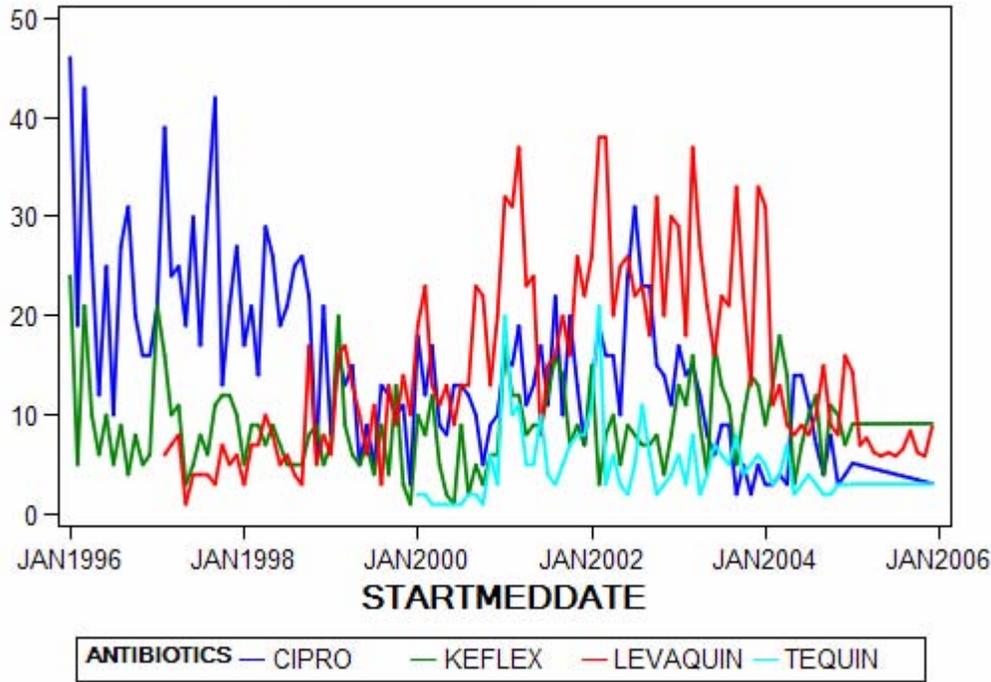


Fig 32. Distribution of Private Insurance for Tequin

The clusters formed based on the severity of the patient condition were compared using Kernel density estimation of the private insurance payment for each antibiotic (Figures 29-31). For patients in cluster 3 taking Tequin (Figure 31), there is a higher payment between 30-115 dollars whereas patients in cluster 6 have a higher probability of paying between 115-180 dollars.

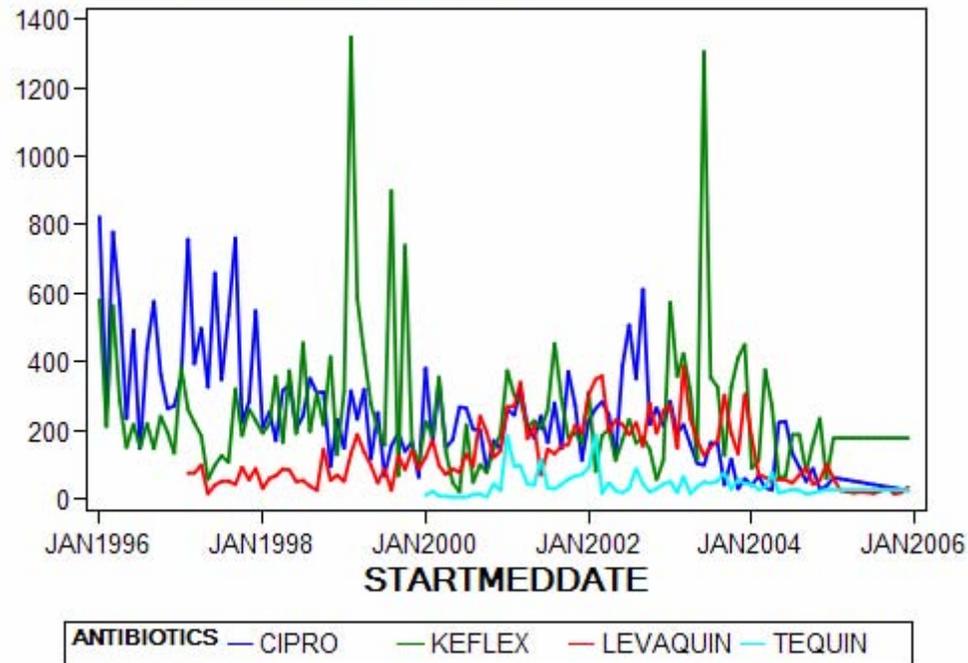
NUMBER OF PRESCRIPTIONS



The number of antibiotics prescribed is increasing starting Jan 2001, with a substantial spike for Levaquin starting Jan 2000.

Figure 33. The number of Antibiotic Prescriptions

QUANTITY OF PRESCRIPTIONS



The quantity of antibiotics prescribed is increasing starting Jan 2001, with a substantial spike for Keflex starting Jan 2003.

Figure 34. The Quantity of Antibiotics Prescriptions

CONCLUSION

High performance forecasting can be applied to healthcare data where the time variable is transactional. Once the trend of the medication is known, researchers can study the medication thoroughly to lower costs and improve care. Severity conditions of a patient were classified using Enterprise Miner (data mining). Probability density estimation was plotted using kernel density estimation methods to compare the total cost and private insurance payment for the severity condition of the patients and it was found that patients with Urinary tract infection have higher frequency. The number of prescription for Levaquin has significantly increased starting Jan-2000 while the number of prescription for Keflex is increasing. Proper use of the data mining and High Performance Forecasting approaches can lower costs and improve the healthcare of society.

REFERENCES

1. Aldrin, M. and Damsleth, E. (1989), "Forecasting Non-seasonal Time Series with Missing Observations," *Journal of Forecasting*, 8, 97-116.
2. <http://www.meps.ahrq.gov/>
3. Anonymous. (2006a). *2006 ICD9-CM and Medical Terminology Dictionary*. ICD9.chrisendres.com. Retrieved, 2006, from the World Wide Web: <http://icd9cm.chrisendres.com/index.php?action=contents>
4. Iazzoni, L. (1997). *Risk adjustment for measuring healthcare outcomes* (2nd ed.) Chicago: Healthcare Administration Press.
5. http://www.medcosm.com/service_icd9lookup.jsp
6. Fraser GL, Stogsdill P, Dickens JD Jr, Wennberg DE, Smith RP Jr, Prato BS. Antibiotic optimization. An evaluation of patient safety and economic outcomes. *Arch Intern Med* 1997; **15**:1689-94.
7. Ansari F, Gray K, Nathwani D, Phillips G, Ogston S, and Ramsay C, et al. Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis. *J Antimicrob Chemother* 2003;**5**:842-8.
8. Chang, M.C. and Dickey, D.A. (1993). "Recognizing Overdifferenced Time Series", *Journal of Time Series Analysis*, **15**,1-18.
9. JHON M. Gottman (1981). *Time series Analysis, A comprehensive Introduction for Social Scientists*. Cambridge University Press.
10. Brockwell, P. J., R.A. Davis, 1996, *Introduction to time series forecasting*, Springer, New York.

AUTHOR CONTACT INFORMATION

Mussie Tesfamicael
Department of Mathematics
University of Louisville
Louisville, KY 40292
Work phone 502-852-6237
Fax 502-852-7132
Email: matesf01@louisville.edu

APPENDIX

SAS Codes

/*This code is used to create an ICD code that will identify the severity of disease*/

```

proc sort data = sasuser.cklt out= work.sort_out;
  by duid rxicd1x;
run;
options obs=max;

data work.sort_out1;
  set work.sort_out;
  icd9 = translate(left(trim(rxicd1x)), '_',' ');
run;

proc Transpose data=work.sort_out1
  out=work.tran /*(drop=_name_ _label_) */
  prefix=icd9_;
  var icd9 ;
  by duid;
run;

data work.concat( keep= duid icd9 ) ;

  length icd9 $32767 ;
  set work.tran ;

  array rxconcat {*} icd9_ ;

  icd9 = left( trim( icd_1 ) ) ;

  do i = 2 to dim( rxconcat ) ;
  icd9 = left(trim(icd9)) || ' ' || left(trim( rxconcat[i] ) ) ;
  end ;
run ;

proc sql ;
  select max( length( icd9 ) ) into :icd9_LEN from work.concat ;
quit ;

%put icd9_LEN=&icd9_LEN ;

data work.concat1 ;
  length icd9 $ &icd9_LEN ;
  set work.concat ;
run ;
proc contents data=work.concat1 ;
run ;

/*Code used to merge ICD code and data with cost and other variables*/
data sasuser.icdcodeswithQuantity (keep= icd9 duid _freq_ Rxpvsu_m_sum);
merge sasuser.Concat1 sasuser.privatesu_m;
by duid;
run;

data PrivateInsurCluster(keep=_Cluster_ _Freq_ Clus_desc duid RXpvsum_sum);
set sasuser.Medicationswithdescriptionscopy2;

```

```
run;
```

```
data sasuser.icdcodeswithTOTALPAY (keep= icd9 duid _freq_ RXXPSUM_Sum);
merge sasuser.Concat1 sasuser.Totalsum;
by duid;
run;
```

SAS Codes

```
/*Kernel Density Probability for Total Payment made*/
ods graphics on;
proc kde data=sasuser.Totalcluster gridl=0 gridu=300 method=srot
out=sasuser.kdeTotalPay;
var RXXPSUM;
by _cluster_;
run;
GOPTIONS RESET=ALL;
GOPTIONS RESET=ALL CTEXT=Bl htext=3.5pct ftext=swissb border;
proc GPLOT data=sasuser.kdeTotalPay ;
plot density*RXXPSUM=_CLUSTER_;
label RXXPSUM="Total Payment";
axis1 label=(a=90 r=0);
symbol1 color=red line=1 w=1 v=none h=1 interpol=join;
symbol2 color=green line=1 w=1 v=none h=1 interpol=join;
symbol3 color=blue line=1 w=1 v=none h=1 interpol=join;
symbol4 color=black line=1 w=1 v=none h=1 interpol=join;
symbol5 color=yellow line=1 w=1 v=none h=1 interpol=join;
symbol6 color=brown line=1 w=1 v=none h=1 interpol=join;
run;
quit;

/*Kernel Density Probability for Private Insurance Payment of a house hold*/
proc sort data=sasuser.Privatecluster;
by _cluster_;
run;
proc kde data=sasuser.Privatecluster gridl=0 gridu=300 method=srot
out=sasuser.kdePrivatePay;
var RXPVSUM;
by _cluster_;
run;
GOPTIONS RESET=ALL;
GOPTIONS RESET=ALL CTEXT=Bl htext=3.5pct ftext=swissb border;
proc GPLOT data=sasuser.kdePrivatePay ;
plot density*RXPVSUM=_CLUSTER_;
label RXPVSUM="Private Insurance Payment";
axis1 label=(a=90 r=0);
symbol1 color=red line=1 w=1 v=none h=1 interpol=join;
symbol2 color=green line=1 w=1 v=none h=1 interpol=join;
symbol3 color=blue line=1 w=1 v=none h=1 interpol=join;
symbol4 color=black line=1 w=1 v=none h=1 interpol=join;
symbol5 color=yellow line=1 w=1 v=none h=1 interpol=join;
symbol6 color=brown line=1 w=1 v=none h=1 interpol=join;
run;
quit;
*****
```